Peta-Scale: Software Challenges beyond 2015

Robert Lupton

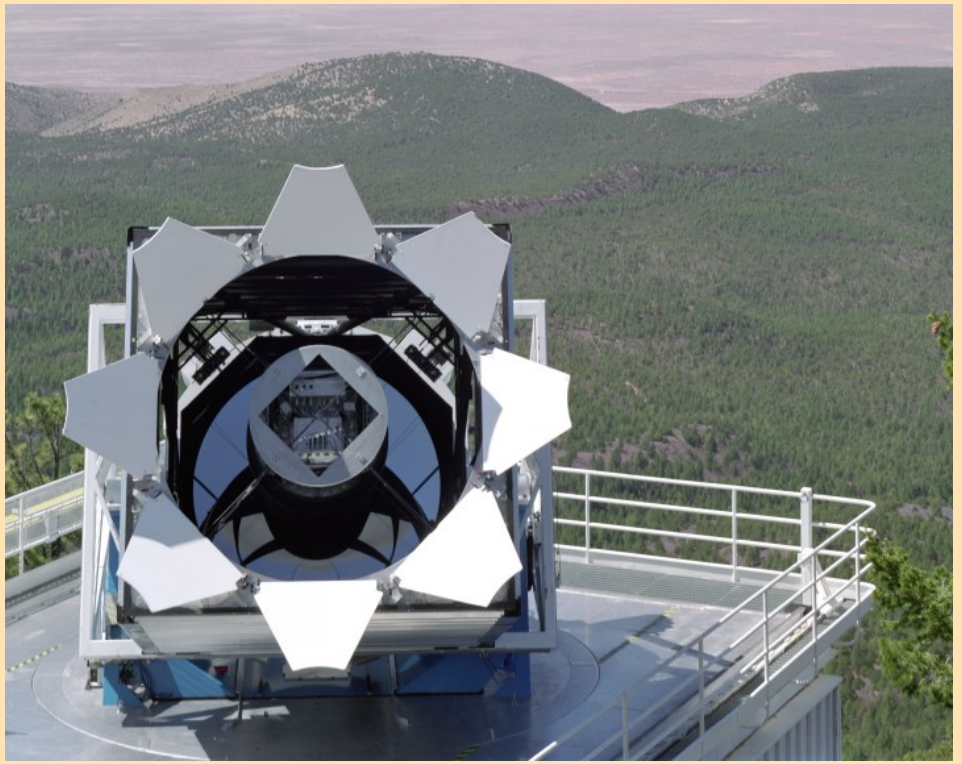
Princeton University

Santa Fe, 21st November, 2008

A Brief Summary of Sloan

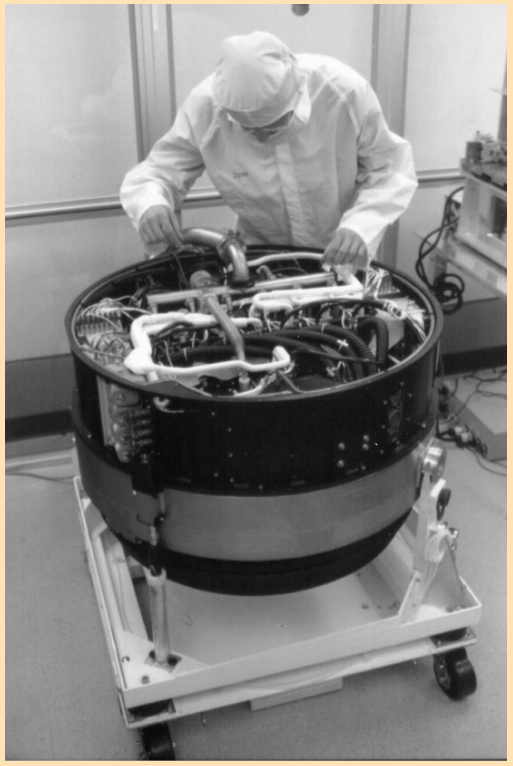
A Brief Summary of Sloan

- A telescope (with a 2.5m diameter primary mirror) at Apache Point, New Mexico



A Brief Summary of Sloan

- Lots of Liquid Nitrogen and Electronics



A Brief Summary of Sloan

- Lots of Liquid Nitrogen and Electronics

A Brief Summary of Sloan

- Lots of Software

```
/*
 * copy the symmetrised image back to the original data region, where
 * it will become the deblending template (the original pixel values
 * are preserved in the parent's atlas image).
 *
 * We must of course only do this within the master_mask
 */
     copy_region_within_mask((REGION *)data, sym, mmask,
                                                  aimage_drow, aimage_dcol);
/*
 * we next want to run the object finder on that symmetrised image; the image
 * is smoothed, and extra peaks rejected --- see improve_template() for details
 */
     obj1 = objc->color[c];
     if(obj1->flags & OBJECT1_DEBLENDED_AS_PSF) {
/*
 * no need to check template, as we created it as a multiple of PSF
 */
     } else {
        float threshold = fiparams->frame[c].ffo_threshold;

        shAssert(obj1->mask == objc->aimage->mask[c]);
        phObjmaskDel(obj1->mask); objc->aimage->mask[c] = NULL;
        obj1->mask =
          improve_template(mmask, c, rowc,colc, data, aimage_drow,aimage_dcol,
                           scra, scrb, rsize + filtsize, csize + filtsize,
                           fiparams->frame[c].smooth_sigma, filtsize,
                           npeak_max, smoothed_ai, threshold, ngrow);

        if(obj1->mask == NULL) {
           objc->flags &= ~OBJECT1_DETECTED;
           obj1->flags &= ~OBJECT1_DETECTED;
        }
     }
  }
/*
 * we've found the templates in all colours. They are represented by the
 * pixels in the original data region, within the OBJECT1->mask
 *
 * Now go through them looking for objects which we didn't detect
 * in any band; in this case, the object wouldn't have been found at all
 * if it wasn't part of a blend, so dump it.
 *
 * Actually we cannot just dump it here as we've got an array with all the
 * children in it, and we'd have to move the others down. Instead, mark
 * the entire OBJC as not DETECTED, and we'll dump it when we get a chance.
 */
  for(c = 0;c < ncolor;c++) {
     objc->flags |= (objc->color[c]->flags & OBJECT1_DETECTED);
  }
  if(!(objc->flags & OBJECT1_DETECTED)) { /* not detected in any band */
     phAtlasImageDel(*smoothed_ai, 0); *smoothed_ai = NULL;
--   deblend.c    38% L1365 CVS-1.128    (C Abbrev)----2:55PM 0.39----------------
```

What's Special about Surveys?

What's Special about Surveys?

Politics

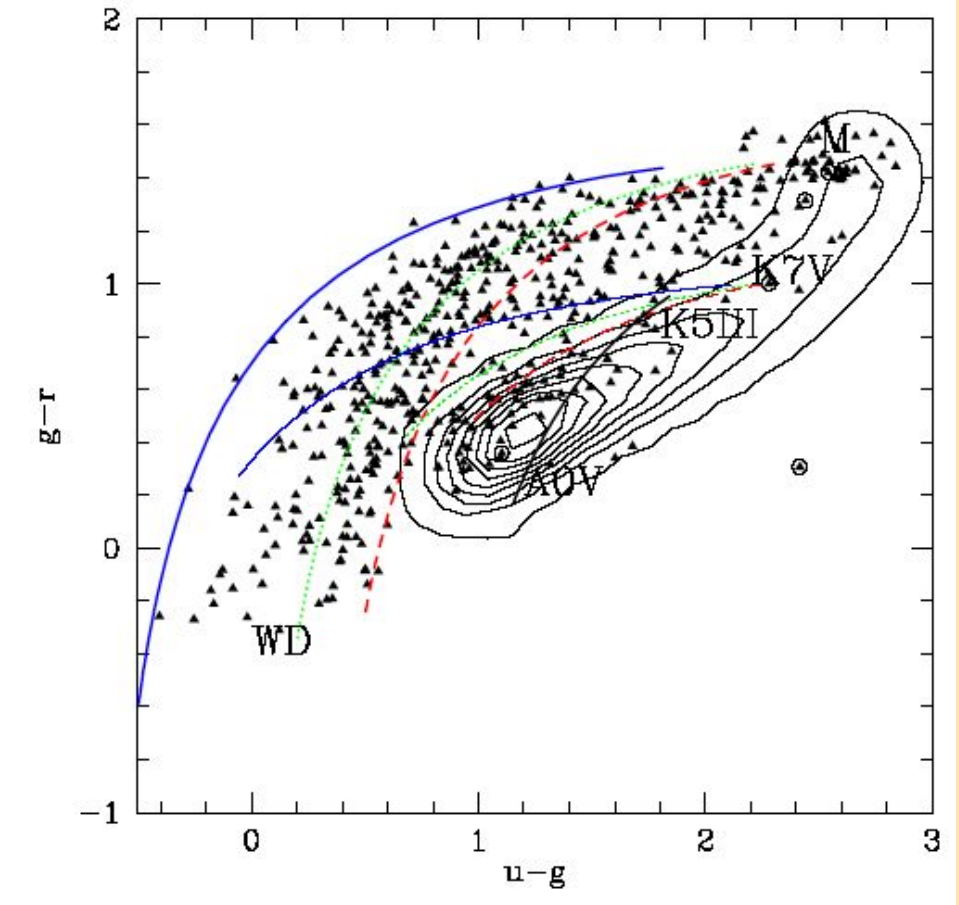
What's Special about Surveys?

Managing Large Collaborations

What's Special about Surveys?

Large Samples

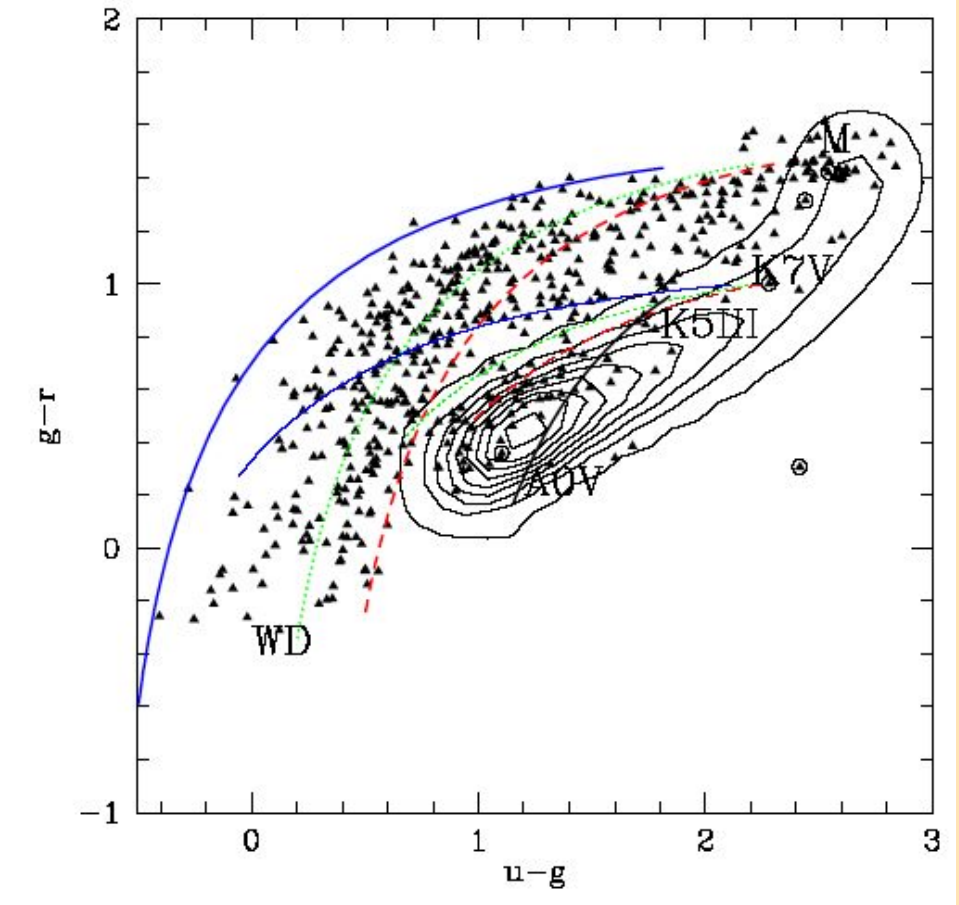
Large Samples



(Pourbaix et al.)

Colour Induced Displacement

Large Samples



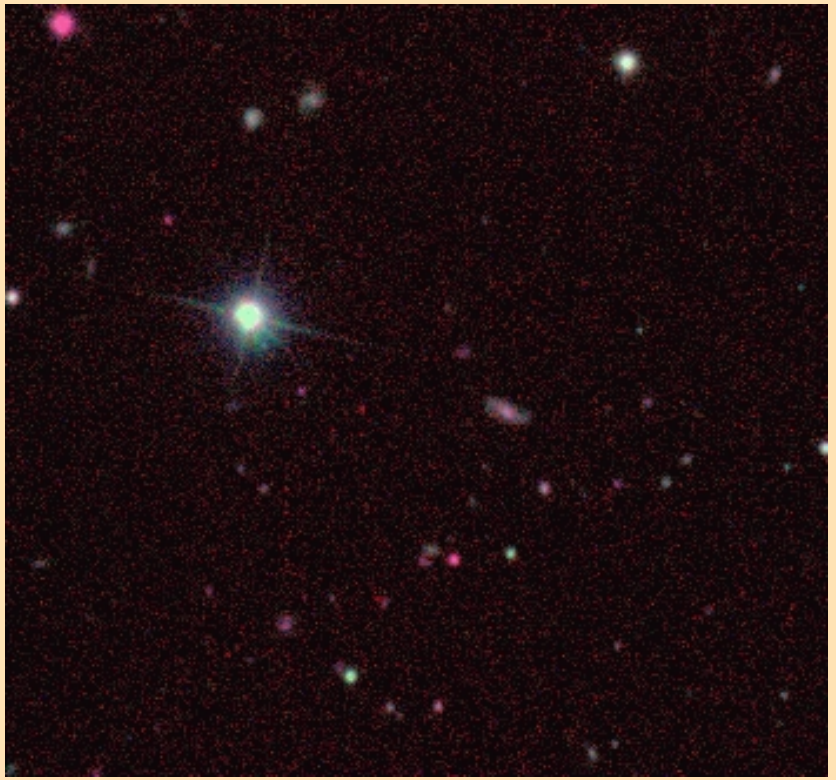
(Pourbaix et al.)

542 binaries out of 5.5×10^6 objects

Rare Objects

Rare Objects

gri



riz

Quality Control

Quality Control

One of the things that we teach our students (and post-docs) is how to look carefully at a set of facts and ask if they makes sense.

Quality Control

One of the things that we teach our students (and post-docs) is how to look carefully at a set of facts and ask if they makes sense.

In the case of theory, this means asking exactly what the New Discovery depends on, and whether its foundations are sound

Quality Control

One of the things that we teach our students (and post-docs) is how to look carefully at a set of facts and ask if they makes sense.

In the case of theory, this means asking exactly what the New Discovery depends on, and whether its foundations are sound

This can mean either the initial assumptions; the methods employed; or the presence of bugs in the code

Quality Control

One of the things that we teach our students (and post-docs) is how to look carefully at a set of facts and ask if they makes sense.

In the case of theory, this means asking exactly what the New Discovery depends on, and whether its foundations are sound

Quality Control

One of the things that we teach our students (and post-docs) is how to look carefully at a set of facts and ask if they makes sense.

In the case of theory, this means asking exactly what the New Discovery depends on, and whether its foundations are sound

In the case of data, this means asking if the Fascinating Result du Jour is an artifact of the instrument or of the reduction.

This is difficult in the context of a survey:

This is difficult in the context of a survey:

- There's too much data for humans to look at

This is difficult in the context of a survey:

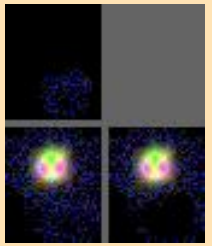
- There's too much data for humans to look at
- The consumer is far removed from the raw data

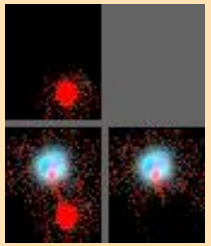
This is difficult in the context of a survey:

- There's too much data for humans to look at
- The consumer is far removed from the raw data
- Large datasets make it possible to study rare events; glitches look like rare events

This is difficult in the context of a survey:

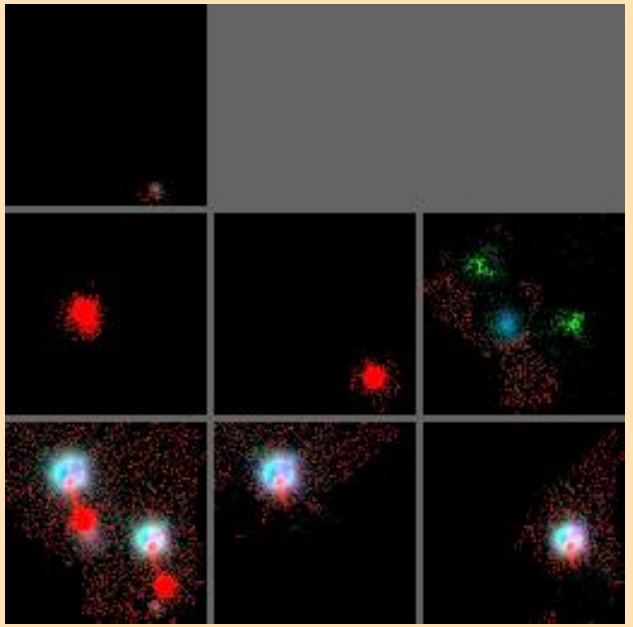
- There's too much data for humans to look at
- The consumer is far removed from the raw data
- Large datasets make it possible to study rare events; glitches look like rare events


ugr


riz

This is difficult in the context of a survey:

- There's too much data for humans to look at
- The consumer is far removed from the raw data
- Large datasets make it possible to study rare events; glitches look like rare events



riz

This is difficult in the context of a survey:

- There's too much data for humans to look at
- The consumer is far removed from the raw data
- Large datasets make it possible to study rare events; glitches look like rare events

This is difficult in the context of a survey:

- There's too much data for humans to look at
- The consumer is far removed from the raw data
- Large datasets make it possible to study rare events; glitches look like rare events

Then there's the problem of how to let the astronomical public what they should trust, and where they should tread warily.

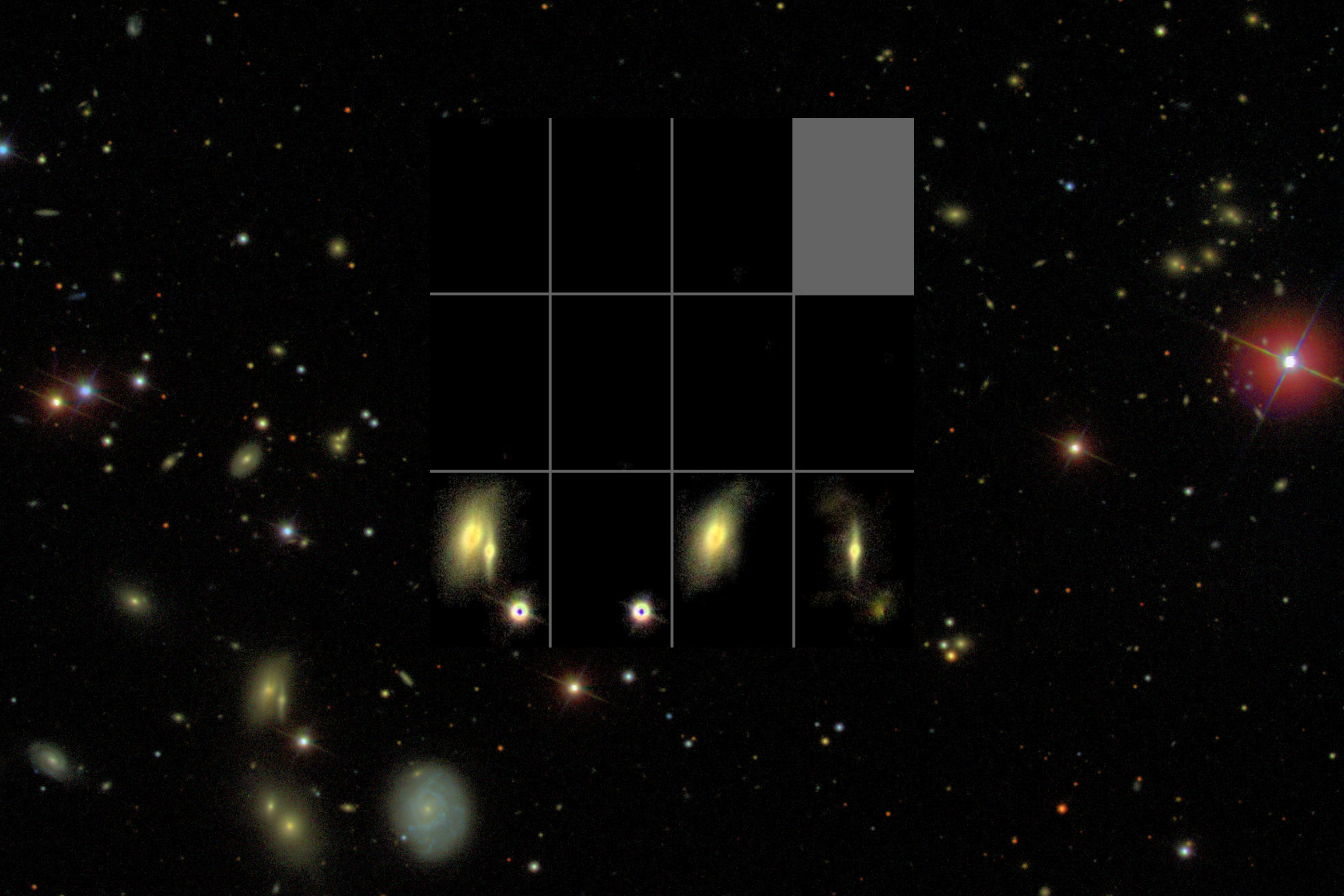
Inside the Sausage Machine: Finding $z \sim 6$ Quasars

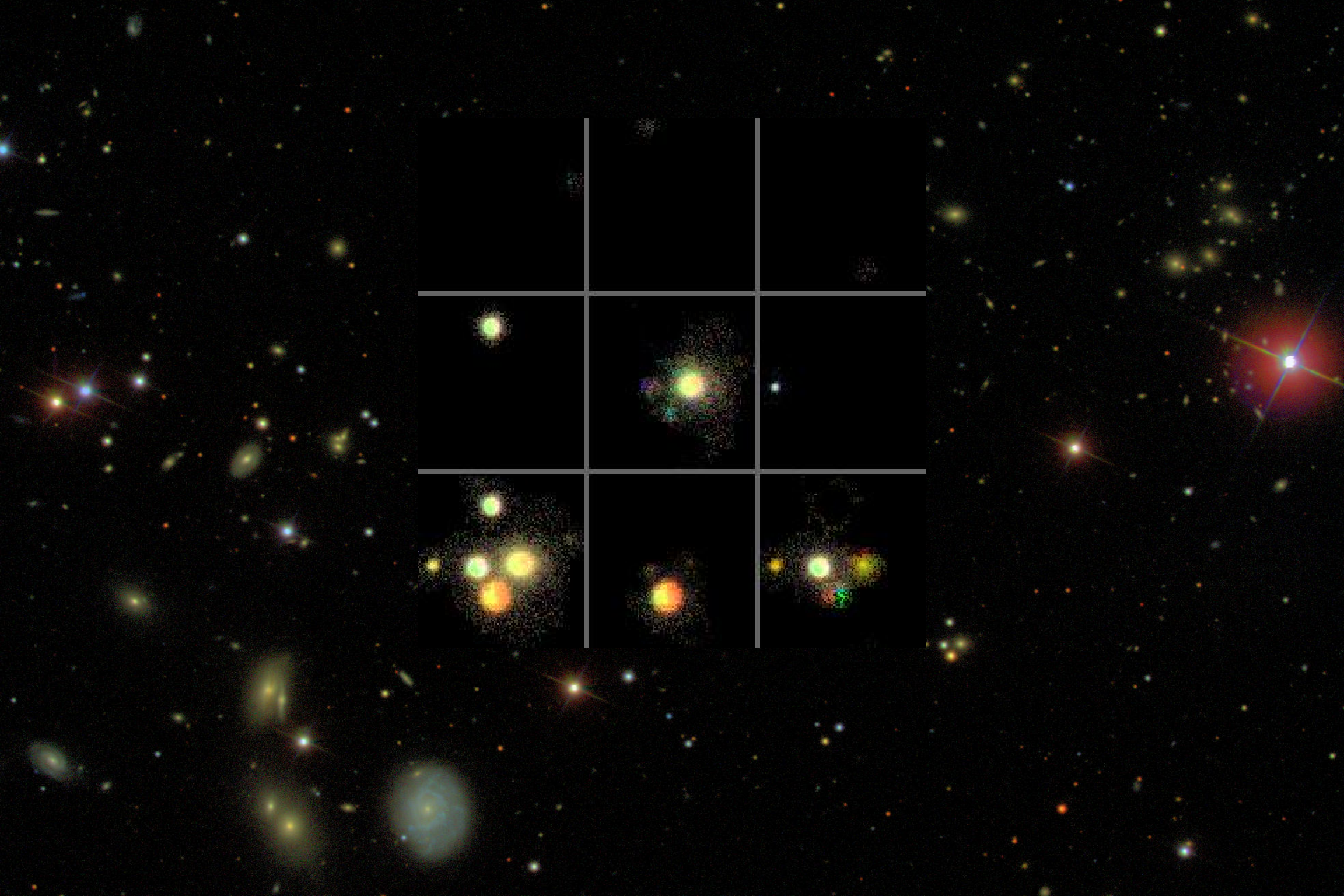
Inside the Sausage Machine: Finding $z \sim 6$ Quasars

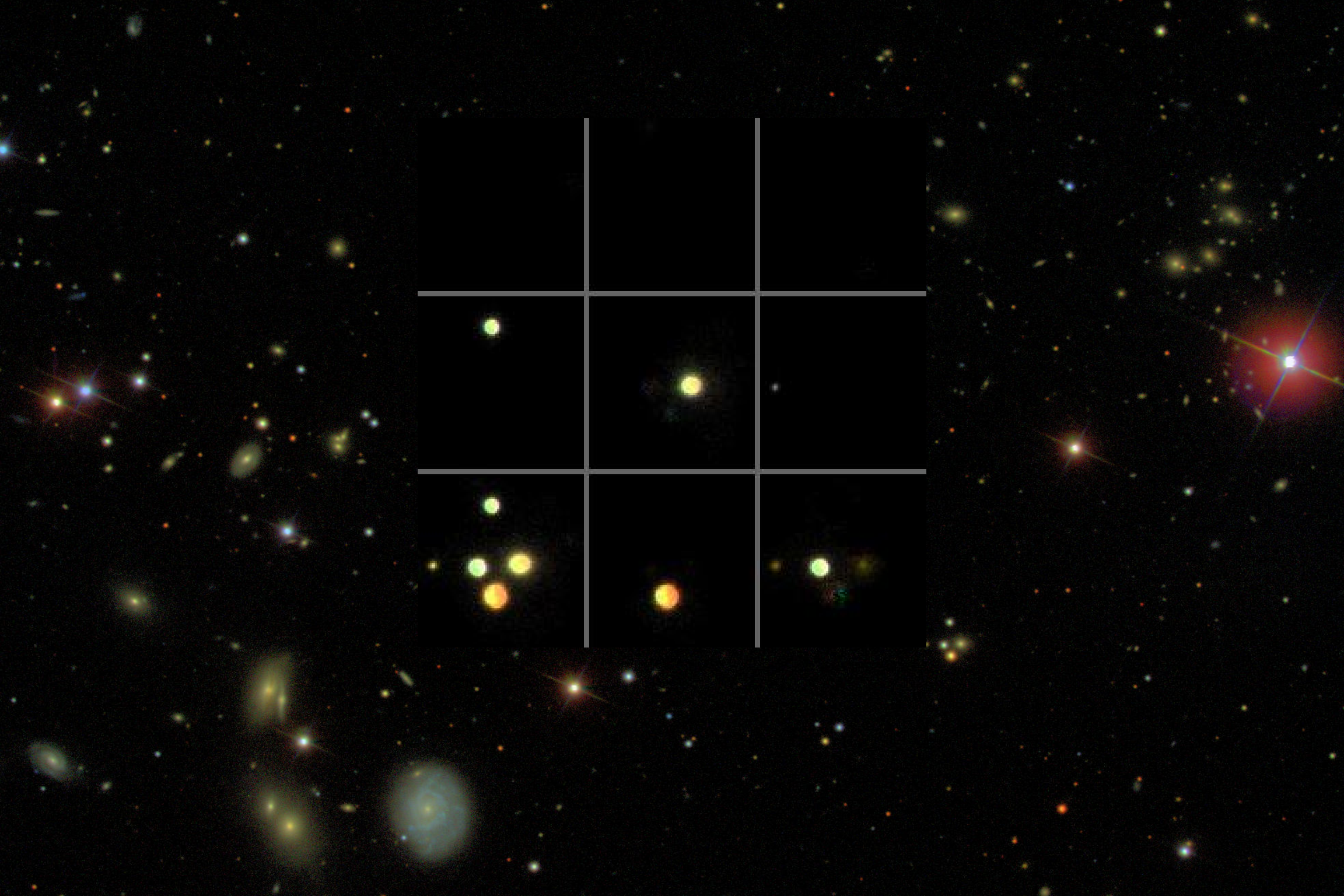
With Xiaohui Fan and Michael Strauss and Željko Ivezić (and ...)

Inside the Sausage Machine: Finding $z \sim 6$ Quasars

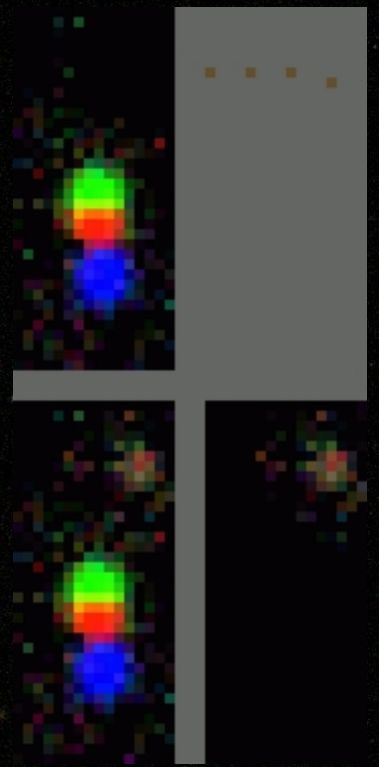

Objects Are Blended

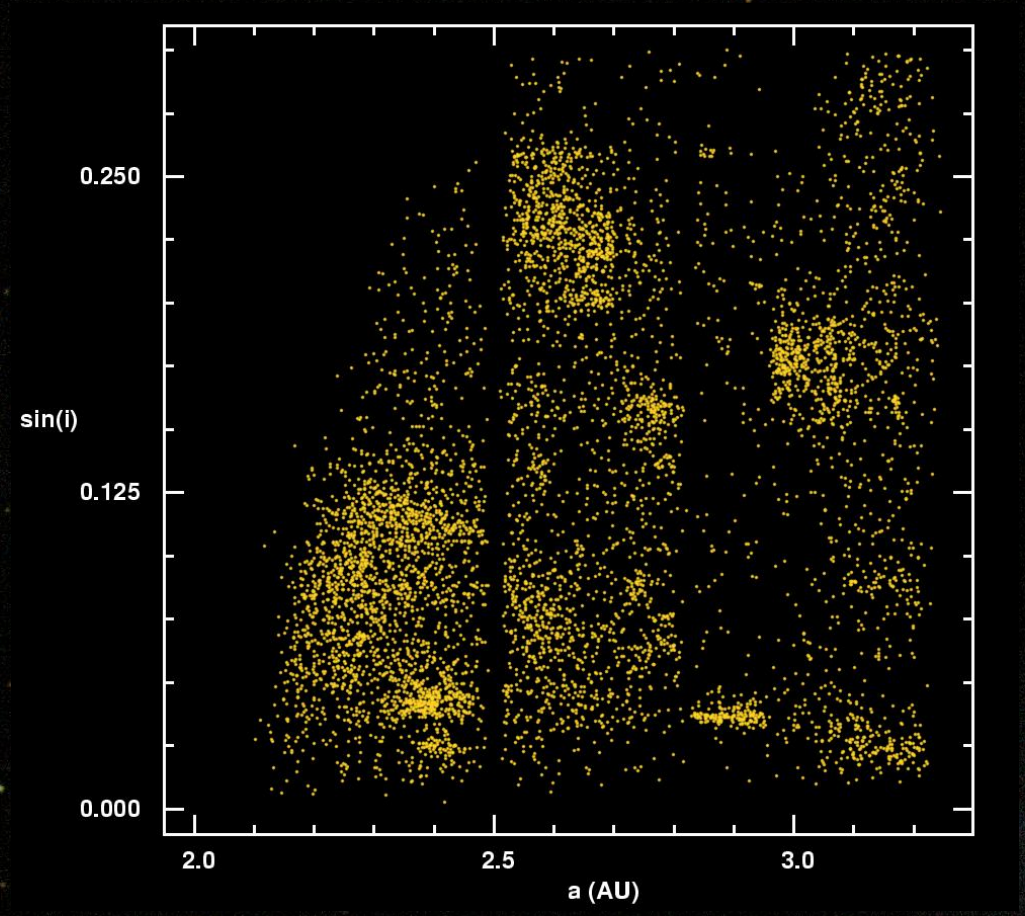




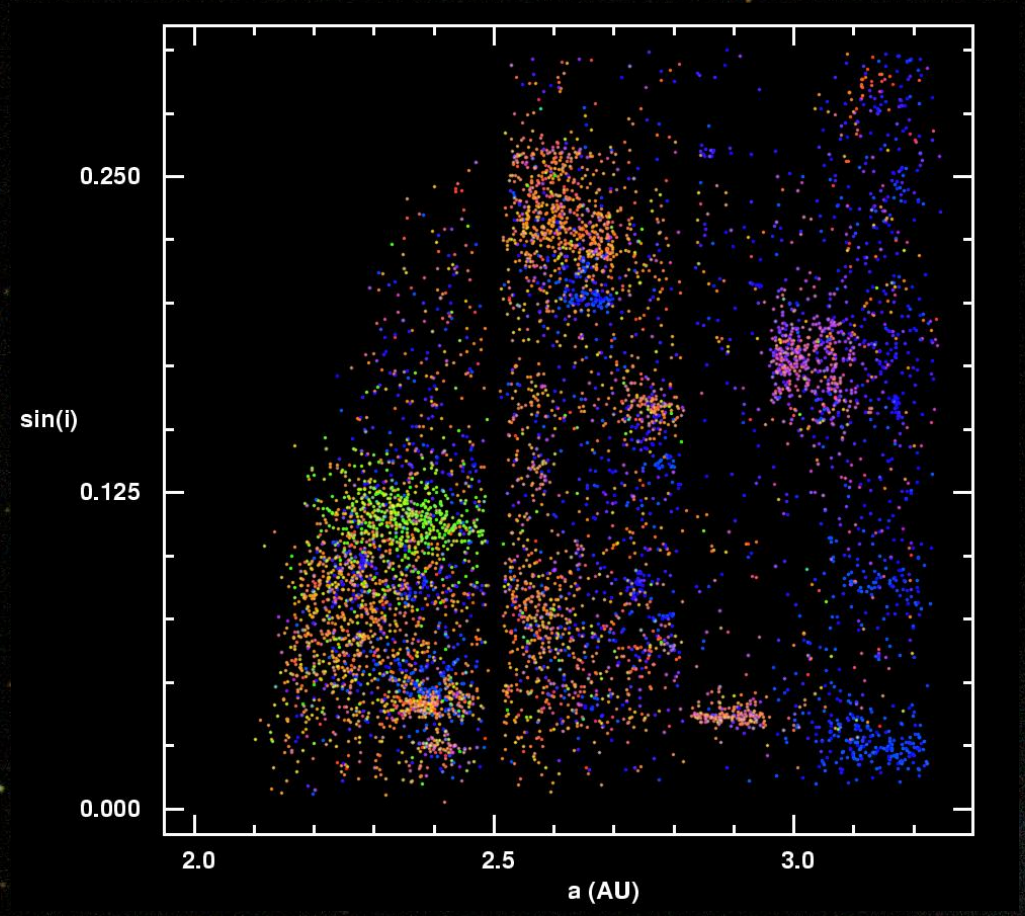


Objects Move





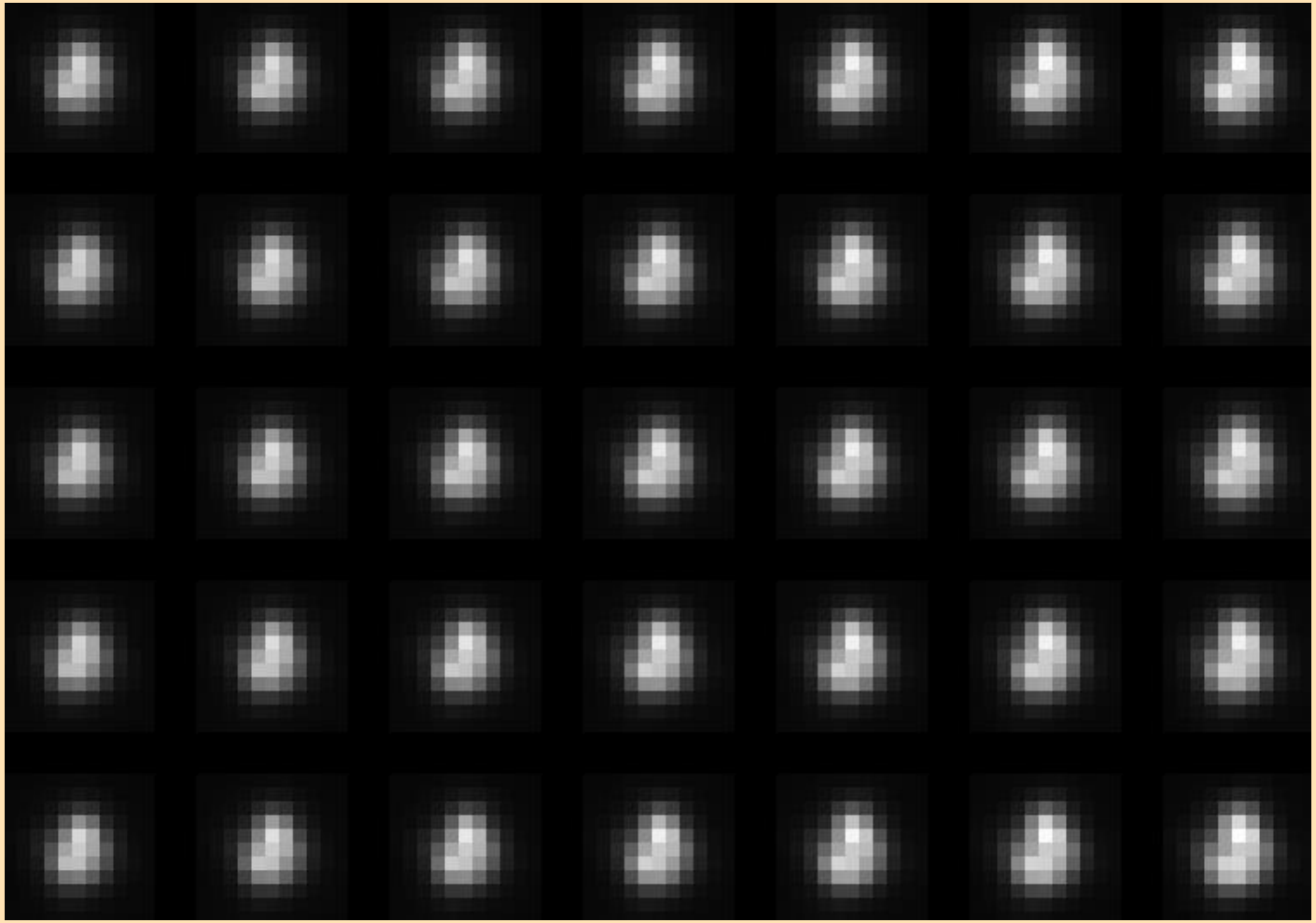
The semi-major axis v. (proper) inclination of a sample of known asteroids detected by SDSS



The semi-major axis v. (proper) inclination of a sample of known asteroids detected by SDSS

The PSF can be Complicated

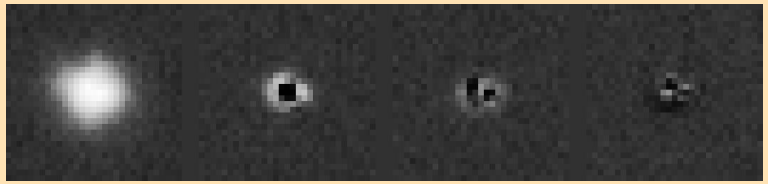
The PSF can be Complicated



The PSF can be Complicated

- KL decompose the bright stars in the frame, giving a number of basis functions (typically 3 or 4):

$$P_{ij} = \sum_{\alpha=0}^{n-1} A^{(\alpha)} K_{ij}^{(\alpha)}$$



- Write the $A^{(\alpha)}$ as low-order polynomials in x, y:

$$P_{ij}(x,y) = \sum_{\alpha=0}^{n-1} \sum_{r=0}^{n_r-1} \sum_{s=0}^{n_s-1} a_j^{(\alpha)} x^r y^s K_{ij}^{(\alpha)}$$

If you combine the last three points:

- blending
- moving
- variable seeing

it is not obvious how to build a catalogue out of a set of observations.

The Next Generation of Imaging Surveys

The Next Generation of Imaging Surveys

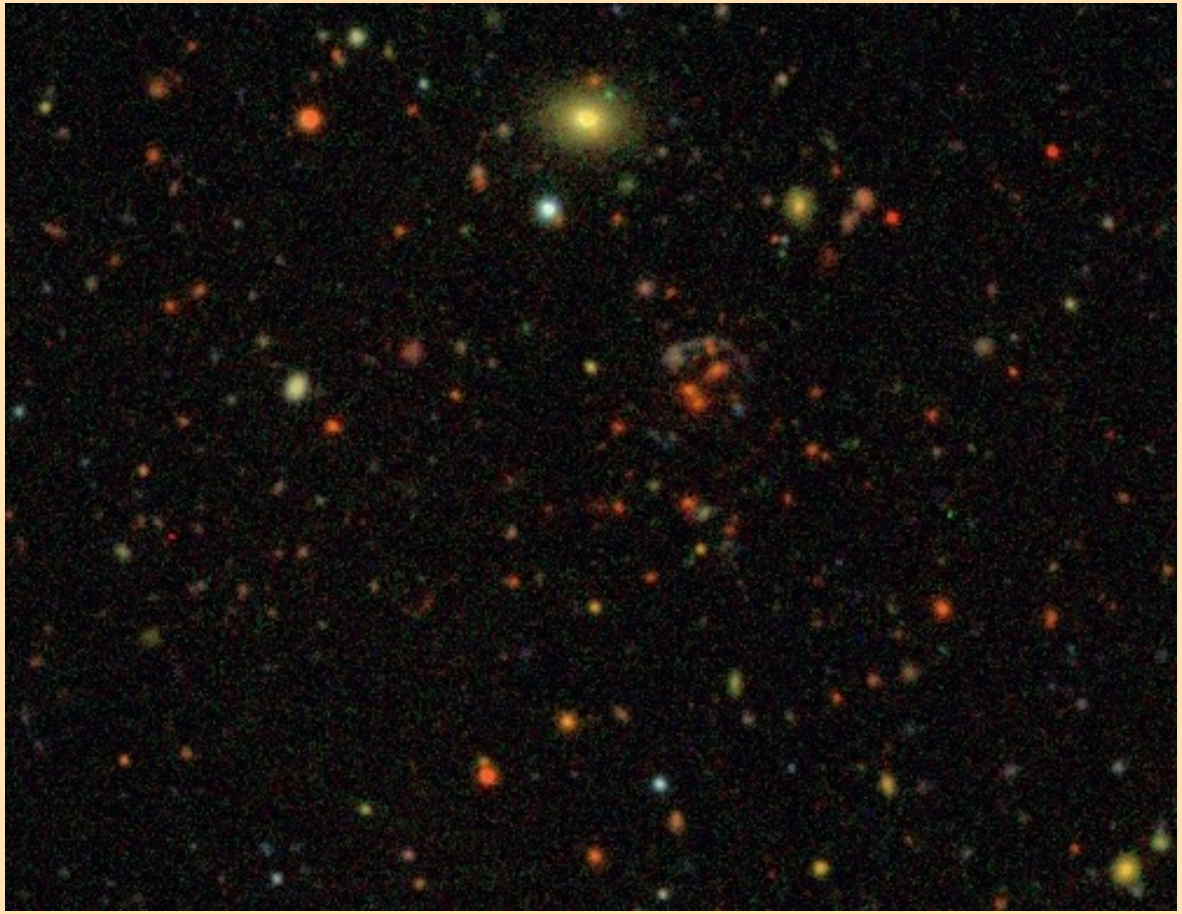
Or, How could you possibly do better than SDSS?

The Next Generation of Imaging Surveys

Or, How could you possibly do better than SDSS?

The Next Generation of Imaging Surveys

Or, How could you possibly do better than SDSS?



The Next Generation of Imaging Surveys

Or, How could you possibly do better than SDSS?

The Next Generation of Imaging Surveys

Or, How could you possibly do better than SDSS?

- More sky coverage

The Next Generation of Imaging Surveys

Or, How could you possibly do better than SDSS?

- More sky coverage
- More epochs

The Next Generation of Imaging Surveys

Or, How could you possibly do better than SDSS?

- More sky coverage
- More epochs
  - Deeper photometry

The Next Generation of Imaging Surveys

Or, How could you possibly do better than SDSS?

- More sky coverage
- More epochs
  - Deeper photometry
  - More reliable photometry

The Next Generation of Imaging Surveys

Or, How could you possibly do better than SDSS?

- More sky coverage
- More epochs
  - Deeper photometry
  - More reliable photometry
  - Variability

The Next Generation of Imaging Surveys

Or, How could you possibly do better than SDSS?

- More sky coverage
- More epochs
  - Deeper photometry
  - More reliable photometry
  - Variability
  - Motions/Parallaxes (1mas/year $\equiv$ 5 km/s at 1 kpc)

The Next Generation of Imaging Surveys

Or, How could you possibly do better than SDSS?

- More sky coverage
- More epochs
  - Deeper photometry
  - More reliable photometry
  - Variability
  - Motions/Parallaxes (1mas/year $\equiv$ 5 km/s at 1 kpc)
- Better image quality

The Next Generation of Imaging Surveys

Or, How could you possibly do better than SDSS?

- More sky coverage
- More epochs
  - Deeper photometry
  - More reliable photometry
  - Variability
  - Motions/Parallaxes (1mas/year $\equiv$ 5 km/s at 1 kpc)
- Better image quality
- More bands (or redder bands)

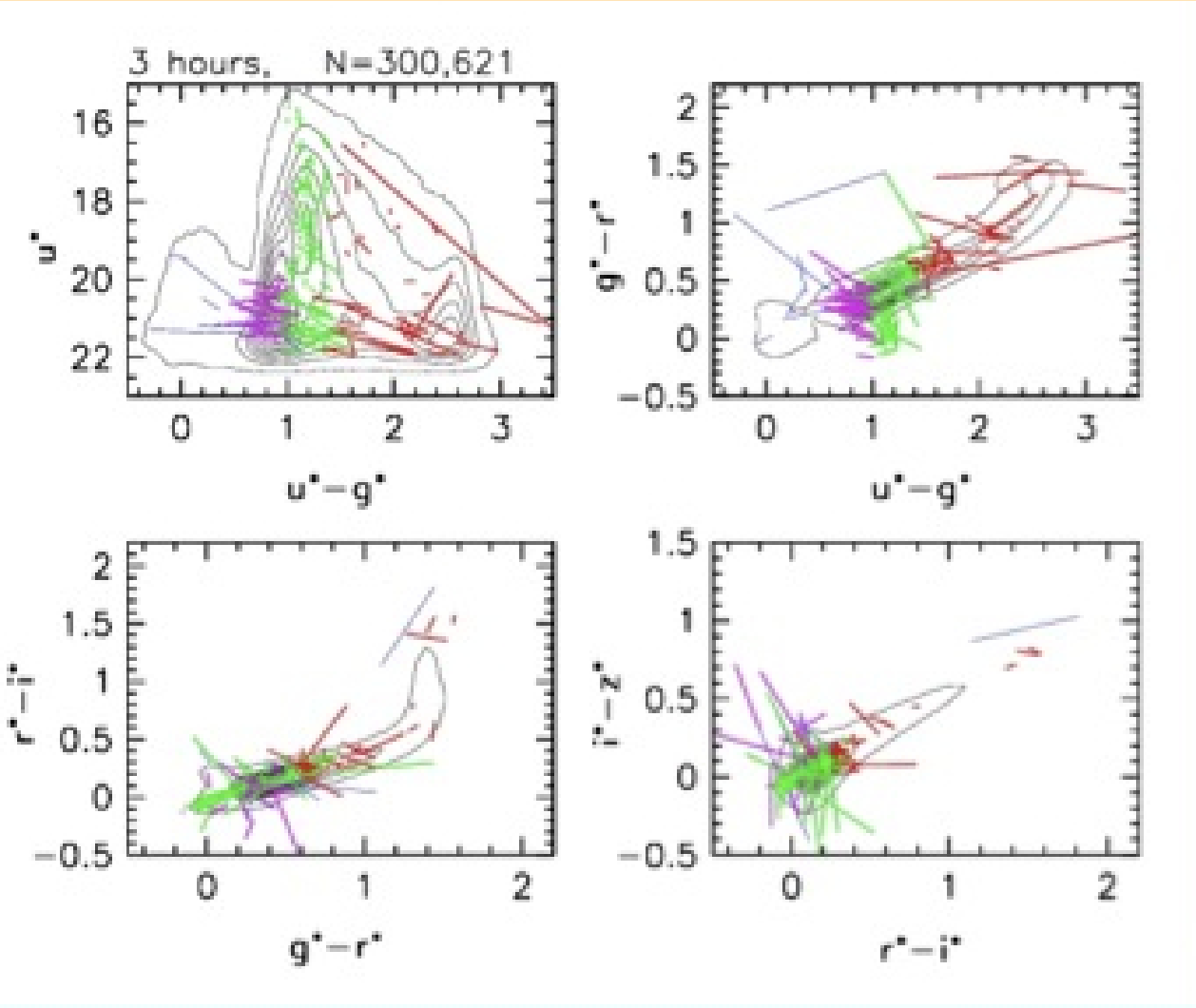
The Next Generation of Imaging Surveys

Or, How could you possibly do better than SDSS?

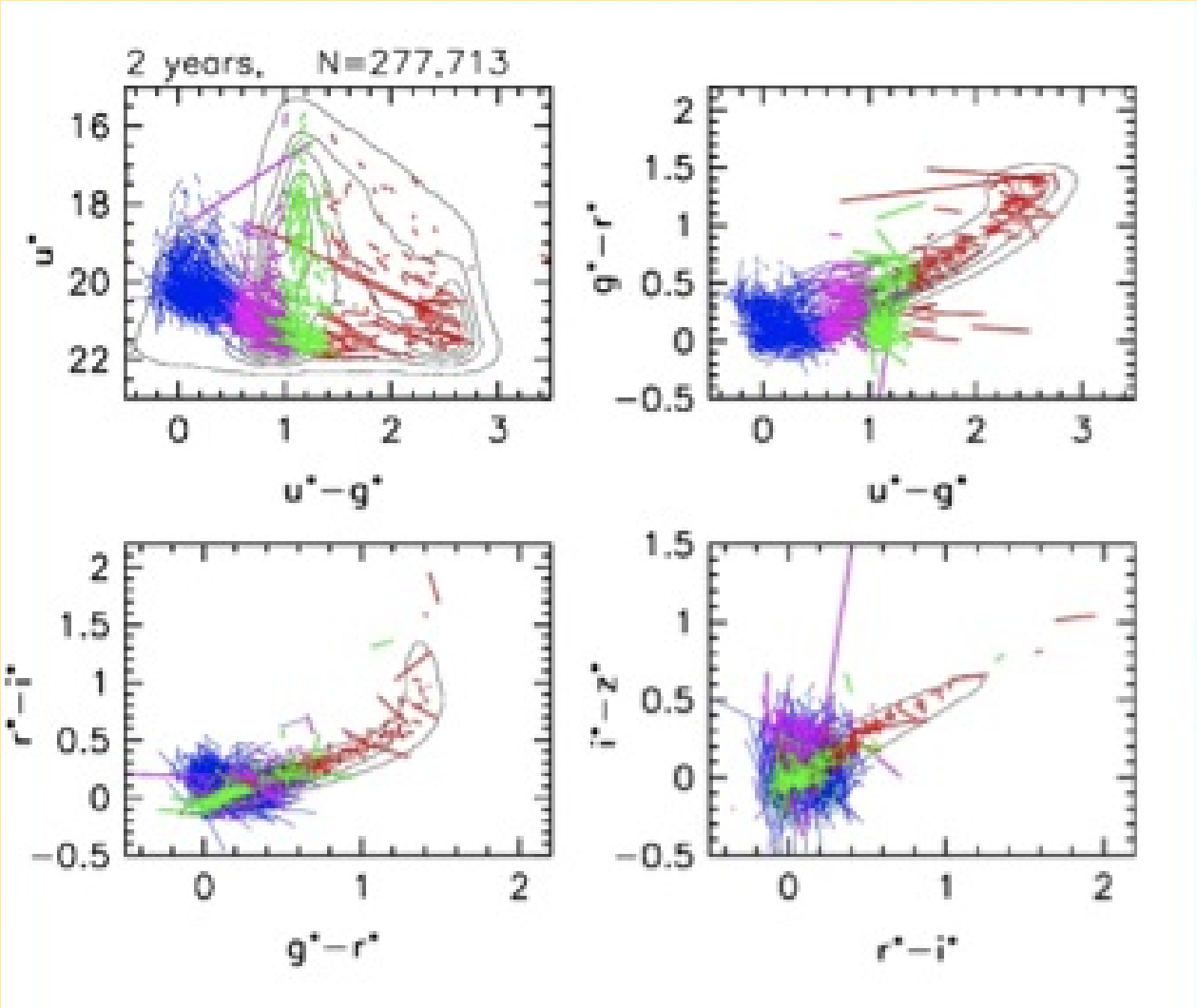
- More sky coverage
- More epochs
  - Deeper photometry
  - More reliable photometry
  - Variability
  - Motions/Parallaxes (1mas/year $\equiv$ 5 km/s at 1 kpc)
- Better image quality
- More bands (or redder bands)
- Better software

E.g. Variability from SDSS

E.g. Variability from SDSS



E.g. Variability from SDSS



What's involved in handling the next generation of data?

What's involved in handling the next generation of data?

- Hardware
  - Disk
  - Processors and Memory. GPUs? Cell Processors?

What's involved in handling the next generation of data?

- Hardware
  - Disk
  - Processors and Memory. GPUs? Cell Processors?

144 16kbit chips

What's involved in handling the next generation of data?

- Hardware
  - Disk
  - Processors and Memory. GPUs? Cell Processors?

What's involved in handling the next generation of data?

- Hardware
  - Disk
  - Processors and Memory. GPUs? Cell Processors?
- Software
  - Algorithms
  - Software Engineering and Techniques
  - Sociology

Software Engineering and Techniques

- Languages (C++ and python?)
- Data types (objects)
- Build systems (or, I hate libtool; LSST uses scons)
- Versioning
- Process management (Naïve ssh? GRID? custom MPI?)
- Fault tolerance
- Provenance
- Testing (regression; science; coverage)
- Data Challenges

Sociology

Sociology

- People

Sociology

- People
- Careers

Sociology

- People
- Careers
- Collaborating at the algorithms level

Sociology

- People
- Careers
- Collaborating at the algorithms level
- Collaborating at the code level

Sociology

- People
- Careers
- Collaborating at the algorithms level
- Collaborating at the code level
- Deciding what's the responsibility of the "Software Group" or the "Scientists"

Sociology

- People
- Careers
- Collaborating at the algorithms level
- Collaborating at the code level
- Deciding what's the responsibility of different Scientists

Processing Polychromatic Sets of Images

Processing Polychromatic Sets of Images

A currently popular approach is to resample the various exposures to a common grid and sum the resulting images with some weighting/filtering. However:

Processing Polychromatic Sets of Images

A currently popular approach is to resample the various exposures to a common grid and sum the resulting images with some weighting/filtering. However:

- Correlated noise

Processing Polychromatic Sets of Images

A currently popular approach is to resample the various exposures to a common grid and sum the resulting images with some weighting/filtering. However:

- Correlated noise
- Sampling

Processing Polychromatic Sets of Images

A currently popular approach is to resample the various exposures to a common grid and sum the resulting images with some weighting/filtering. However:

- Correlated noise
- Sampling
- Discontinuous PSFs

Processing Polychromatic Sets of Images

A currently popular approach is to resample the various exposures to a common grid and sum the resulting images with some weighting/filtering. However:

- Correlated noise
- Sampling
- Discontinuous PSFs
- No opportunity for non-linear analysis in the processing (e.g. 3σ clips).

Processing Polychromatic Sets of Images

A currently popular approach is to resample the various exposures to a common grid and sum the resulting images with some weighting/filtering. However:

- Correlated noise
- Sampling
- Discontinuous PSFs
- No opportunity for non-linear analysis in the processing (e.g. 3σ clips).
- Average over moving/variable objects

Processing Polychromatic Sets of Images

A currently popular approach is to resample the various exposures to a common grid and sum the resulting images with some weighting/filtering. However:

- Correlated noise
- Sampling
- Discontinuous PSFs
- No opportunity for non-linear analysis in the processing (e.g. 3σ clips).
- Average over moving/variable objects

On the other-hand, it has the great advantage of being computationally relatively simple and cheap.

An easy alternative is to process each exposure separately, and add the resulting measurements.

An easy alternative is to process each exposure separately, and add the resulting measurements.

- Only objects detected in at least one frame are measured

An easy alternative is to process each exposure separately, and add the resulting measurements.

- Only objects detected in at least one frame are measured
- There is no guarantee that the same objects will be detected in each exposure

An easy alternative is to process each exposure separately, and add the resulting measurements.

- Only objects detected in at least one frame are measured
- There is no guarantee that the same objects will be detected in each exposure
- It seems unlikely that the errors in all measurements (e.g. galaxy effective radii) will scale as $\sqrt{N}$.

An easy alternative is to process each exposure separately, and add the resulting measurements.

- Only objects detected in at least one frame are measured
- There is no guarantee that the same objects will be detected in each exposure
- It seems unlikely that the errors in all measurements (e.g. galaxy effective radii) will scale as $\sqrt{N}$.

There are ways around some of these problems; for example, we could *detect* on a coadded frame and then use this master catalogue to measure each of the input images.

A new generation of analysis codes should:

A new generation of analysis codes should:

- Never resample the data

A new generation of analysis codes should:

- Never resample the data
- Analyse stacks of data (taken in multiple bands) as a series of samples of the

A new generation of analysis codes should:

- Never resample the data
- Analyse stacks of data (taken in multiple bands) as a series of samples of the

  sky, rather than attempt to generate a single image.
- Make full use of the per-exposure PSF information

A new generation of analysis codes should:

- Never resample the data
- Analyse stacks of data (taken in multiple bands) as a series of samples of the

  sky, rather than attempt to generate a single image.
- Make full use of the per-exposure PSF information
- Preserve variability information (astrometric and photometric)

A new generation of analysis codes should:

- Never resample the data
- Analyse stacks of data (taken in multiple bands) as a series of samples of the

  sky, rather than attempt to generate a single image.
- Make full use of the per-exposure PSF information
- Preserve variability information (astrometric and photometric)
- Use some standard software framework

(Semi-?) Open algorithmic questions

(Semi-?) Open algorithmic questions

- Estimating the PSF and its spatial structure

(Semi-?) Open algorithmic questions

- Estimating the PSF and its spatial structure
- Detecting objects (resolved/trailed; χ^2 image or given SED or ...)

(Semi-?) Open algorithmic questions

- Estimating the PSF and its spatial structure
- Detecting objects (resolved/trailed; χ^2 image or given SED or ...)
- Deblending of stars and galaxies

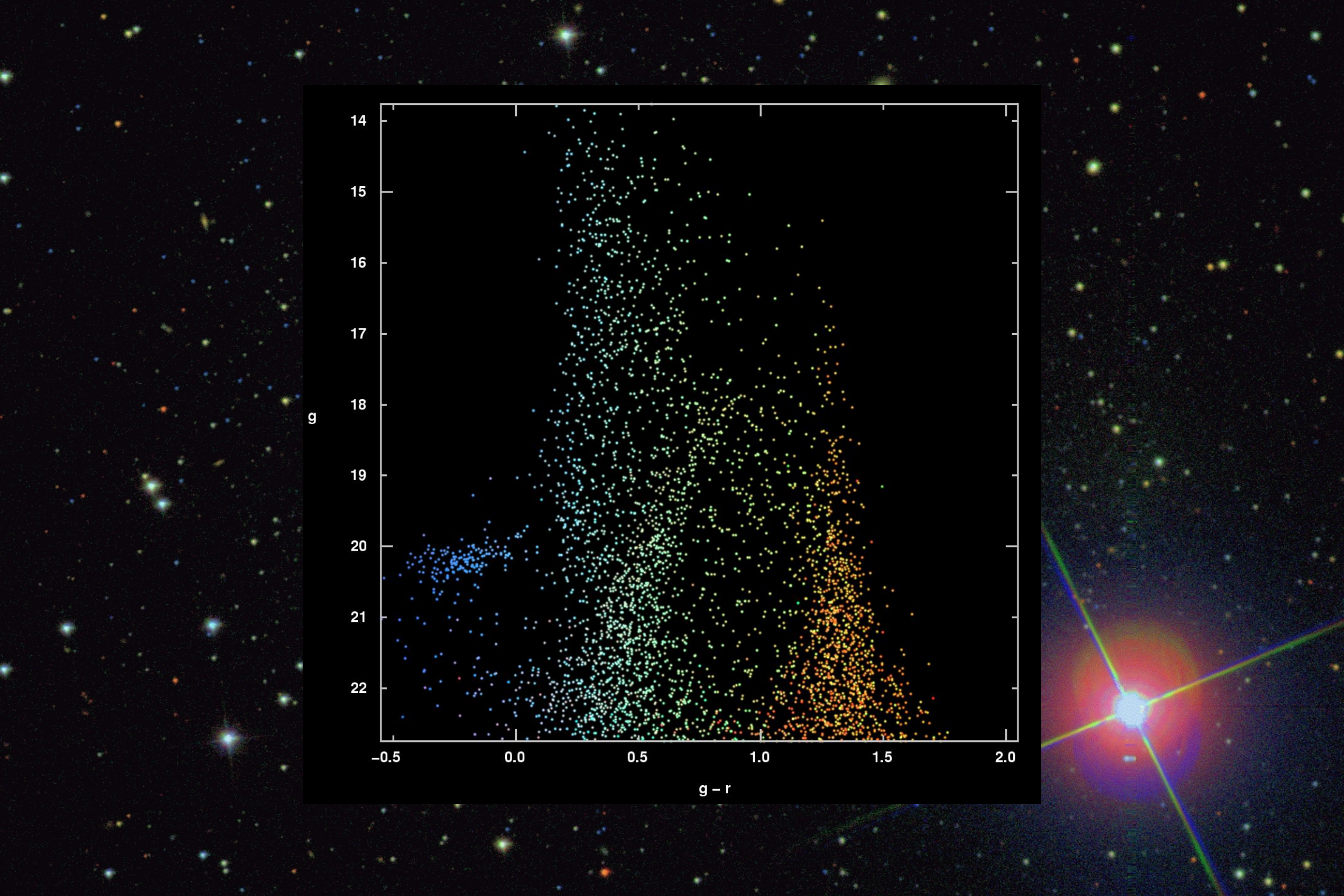
(Semi-?) Open algorithmic questions

- Estimating the PSF and its spatial structure
- Detecting objects (resolved/trailed; χ^2 image or given SED or ...)
- Deblending of stars and galaxies
- Shape measurements

(Semi-?) Open algorithmic questions

- Estimating the PSF and its spatial structure
- Detecting objects (resolved/trailed; χ^2 image or given SED or ...)
- Deblending of stars and galaxies
- Shape measurements
- (Galaxy) photometry


The End


14
15
16
17
18
19
20
21
22
g
−0.5
0.0
0.5
1.0
1.5
2.0
g − r

How should I coadd a set of images?

How should I coadd a set of images?

Caveat: I stole some of these ideas from Nick Kaiser

How should I coadd a set of images?

Given a set of images of the same part of the sky, how should I add them to obtain a deeper image?

How should I coadd a set of images?

Given a set of images of the same part of the sky, how should I add them to obtain a deeper image?

- How far does $\sqrt{N}$ take you?

How should I coadd a set of images?

Given a set of images of the same part of the sky, how should I add them to obtain a deeper image?

- How far does $\sqrt{N}$ take you?
- What's a good algorithm?

How should I coadd a set of images?

Given a set of images of the same part of the sky, how should I add them to obtain a deeper image?

- How far does $\sqrt{N}$ take you?
- What's a good algorithm?
- Is there an optimal algorithm?

How should I coadd a set of images?

Given a set of images of the same part of the sky, how should I add them to obtain a deeper image?

- How far does $\sqrt{N}$ take you?
- What's a good algorithm?
- Is there an optimal algorithm?

There are (at least) three ways to think about adding images:

How should I coadd a set of images?

Given a set of images of the same part of the sky, how should I add them to obtain a deeper image?

- How far does $\sqrt{N}$ take you?
- What's a good algorithm?
- Is there an optimal algorithm?

There are (at least) three ways to think about adding images:

- Add the images together

How should I coadd a set of images?

Given a set of images of the same part of the sky, how should I add them to obtain a deeper image?

- How far does $\sqrt{N}$ take you?
- What's a good algorithm?
- Is there an optimal algorithm?

There are (at least) three ways to think about adding images:

- Add the images together
- Estimate a picture of the Universe

How should I coadd a set of images?

Given a set of images of the same part of the sky, how should I add them to obtain a deeper image?

- How far does $\sqrt{N}$ take you?
- What's a good algorithm?
- Is there an optimal algorithm?

There are (at least) three ways to think about adding images:

- Add the images together
- Estimate a picture of the Universe
- Estimate the properties of the Universe

Estimating a Picture of the Universe

Estimating a Picture of the Universe

If we take the middle tack, we can write down the ML estimate of the Universe U given an image, I, and a (known) PSF, ϕ:

Estimating a Picture of the Universe

If we take the middle tack, we can write down the ML estimate of the Universe U given an image, I, and a (known) PSF, ϕ:

$$I(x) = U(x) \otimes \phi(x) + \epsilon(x)$$

Estimating a Picture of the Universe

If we take the middle tack, we can write down the ML estimate of the Universe U given an image, I, and a (known) PSF, ϕ:

$$I(k) = U(k) \times \phi(k) + \epsilon(k)$$

Estimating a Picture of the Universe

If we take the middle tack, we can write down the ML estimate of the Universe U given an image, I, and a (known) PSF, ϕ:

$$I(k) = U(k) \times \phi(k) + \epsilon(k)$$

Let us assume that all objects are fainter than the sky, so ϵ is an $N(0, \sigma^2)$ variate.

Estimating a Picture of the Universe

If we take the middle tack, we can write down the ML estimate of the Universe U given an image, I, and a (known) PSF, ϕ:

$$I(k) = U(k) \times \phi(k) + \epsilon(k)$$

Let us assume that all objects are fainter than the sky, so ϵ is an $N(0, \sigma^2)$ variate.

$$\ln \mathcal{L} \propto -\ln \sigma - \tfrac{1}{2} \left(U\phi - I\right)^2 / \sigma^2$$

Estimating a Picture of the Universe

If we take the middle tack, we can write down the ML estimate of the Universe U given an image, I, and a (known) PSF, ϕ:

$$I(k) = U(k) \times \phi(k) + \epsilon(k)$$

Let us assume that all objects are fainter than the sky, so ϵ is an $N(0, \sigma^2)$ variate.

$$\ln \mathcal{L} \propto -\sum_i \ln \sigma_i - \tfrac{1}{2} \sum_i \left(U\phi_i - I_i\right)^2 / \sigma_i^2$$

Estimating a Picture of the Universe

If we take the middle tack, we can write down the ML estimate of the Universe U given an image, I, and a (known) PSF, ϕ:

$$I(k) = U(k) \times \phi(k) + \epsilon(k)$$

Let us assume that all objects are fainter than the sky, so ϵ is an $N(0, \sigma^2)$ variate.

$$\ln \mathcal{L} \propto -\sum_i \ln \sigma_i - \tfrac{1}{2} \sum_i \left(U\phi_i - I_i\right)^2 / \sigma_i^2$$

so, differentiating with respect to the Universe,

$$U(k) = \frac{\sum_i I_i \phi_i / \sigma_i^2}{\sum_i \phi_i^2 / \sigma_i^2} \equiv \frac{D(k)}{P(k)}$$

An Optimal Algorithm

An Optimal Algorithm

$$U(k) = \frac{D(k)}{P(k)}$$

$$D(k) \equiv \sum_i I_i \phi_i / \sigma_i^2; \qquad P(k) \equiv \sum_i \phi_i^2 / \sigma_i^2$$

An Optimal Algorithm

$$U(k) = \frac{D(k)}{P(k)}$$

$$D(k) \equiv \sum_i I_i \phi_i / \sigma_i^2; \qquad P(k) \equiv \sum_i \phi_i^2 / \sigma_i^2$$

I.e.

$$U(x) = D(x) \otimes^{-1} P(x)$$

where

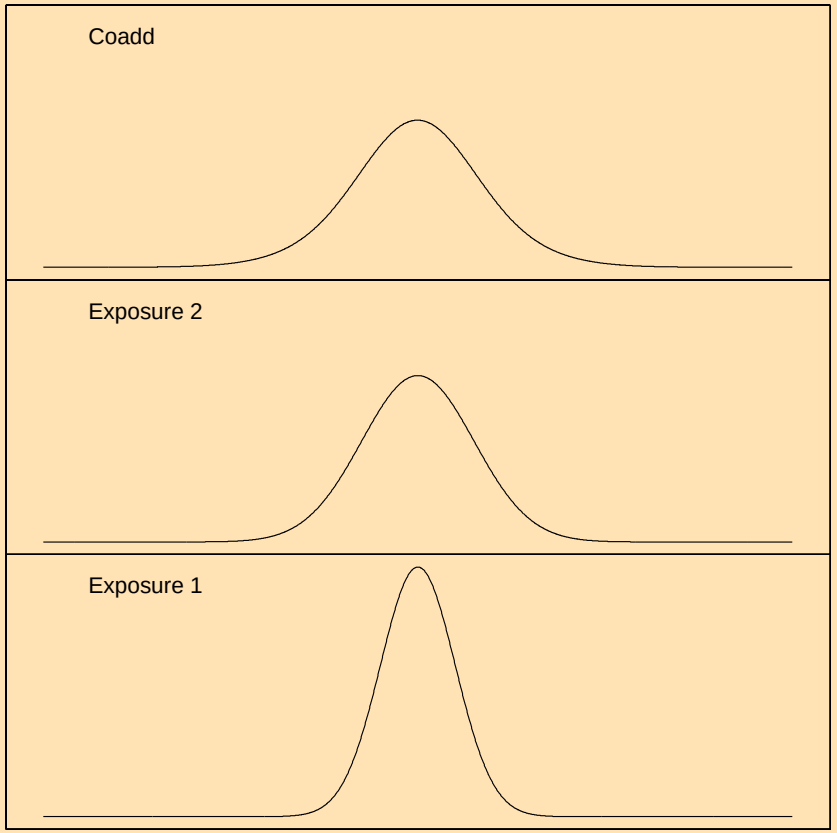
$$D(x) = \sum_i I_i \otimes \phi_i / \sigma_i^2; \qquad P(x) = \sum_i \phi_i \otimes \phi_i / \sigma_i^2$$

Is this Wise?

Probably not.

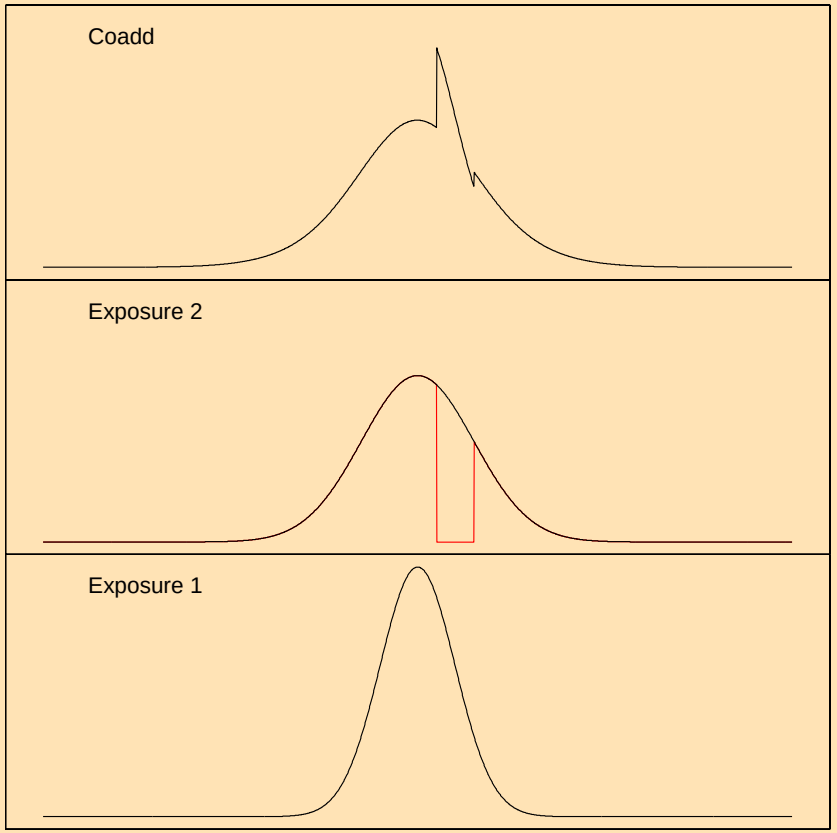
Is this Wise?

Probably not.



Is this Wise?

Probably not.



Estimate the properties of the Universe

This is straightforward for e.g. PSF magnitudes.

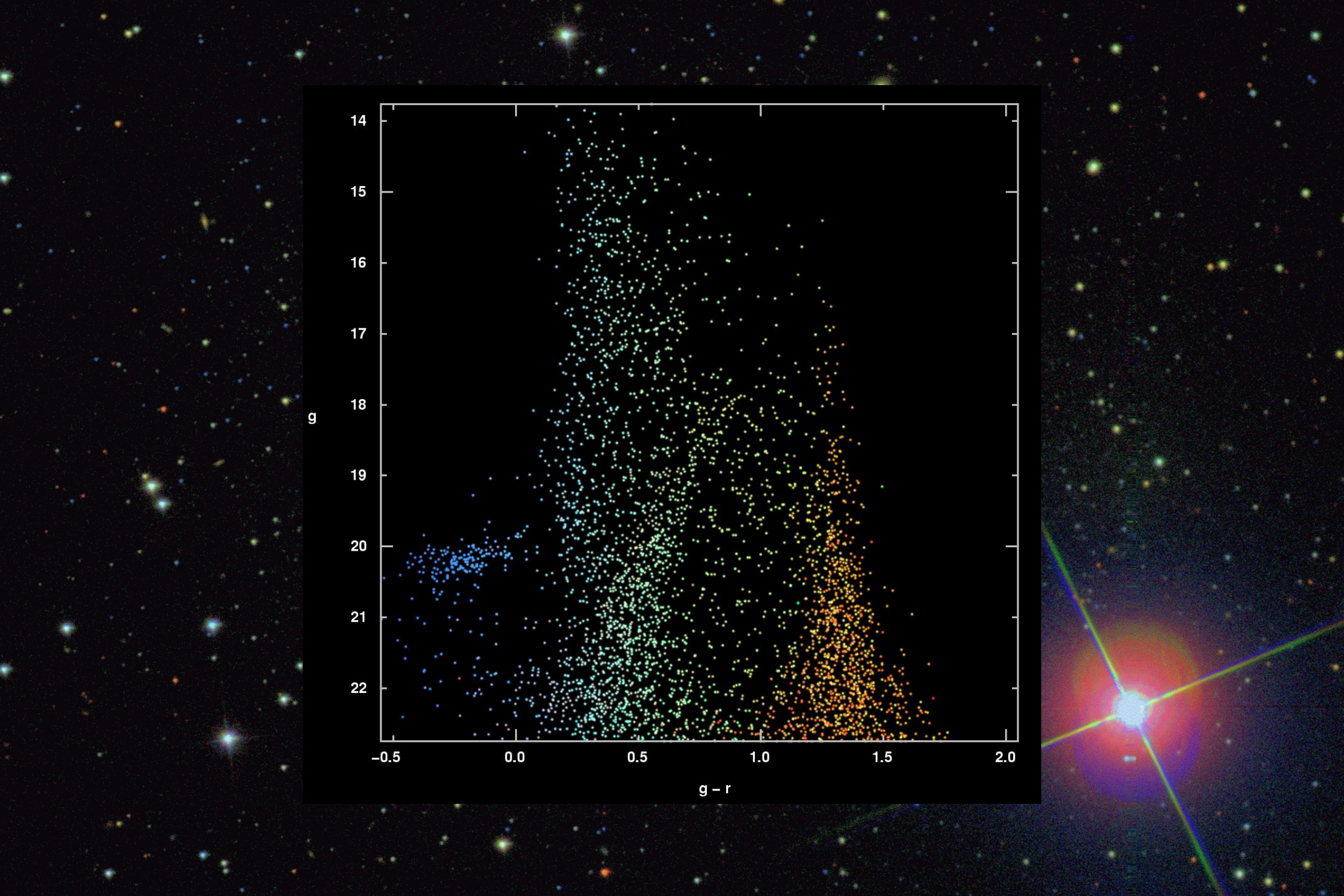
Estimate the properties of the Universe

This is straightforward for e.g. PSF magnitudes.

Harder problems include:

- Sky estimation
- Object detection
- Deblending
- Shape measurements


The End


14
15
16
17
18
19
20
21
22
g
−0.5
0.0
0.5
1.0
1.5
2.0
g − r